# Exploratory Analysis, what it should be and what it should not be

## I recently had a collaborator try to justify doing far too many analyses on a small dataset by claiming they were only exploratory, and they did not have a specific hypothesis to test.

## Nevertheless, they wanted to generate p-values testing dozens of potential hypotheses. In my mind this is more of a fishing expedition than an “exploratory” analysis

# I am very doubtful that, given the pressure to write papers

# What exploratory analysis should be

## Sometimes the best thing to do is generate data driven hypotheses

## Usually it is better science to do in depth descriptive data exploration

### Plots

### Tables

### Distributions

## Don’t make hypotheses based on data you don’t have.

# Data-driven approaches for generating hypotheses

## This is machine learning.

# Large Scale Inference (Empirical Bayes)

## Genetics research by nature involves testing thousands of hypotheses. This is the ideal example of data driven hypothesis generation.

# What is the question: which x predicts y OR does x predict y

# <https://www.epa.gov/caddis/exploratory-data-analysis#:~:text=Exploratory%20Data%20Analysis%20(EDA)%20is,step%20in%20any%20data%20analysis>.

# <https://www.ibm.com/topics/exploratory-data-analysis>

# <https://www.tandfonline.com/doi/full/10.1080/09515089.2022.2113771>

# <https://www.sciencedirect.com/science/article/pii/S1053482216300365?casa_token=z6PQaIpEBC8AAAAA:lZlC0mTS-jAdgu02ch3jVqUdjxhkDeTidf_Snyao5g5BtImDgZmdviY_tKv172fwoDUls4c>

# <https://www.mdpi.com/2072-6643/8/10/593>

# <https://www.nature.com/articles/nmeth.3885>